Women's Commission Report on Teen Pregnancy

Jan Nelson Schrol
Director
Wyoming Commission on the status of Women
Hathaway Building
Cheyenne, WY 82002
(307) 777-7349

Young Parents Program

Ms. Jean Rustici
Consultant, Early Childhood, Parenting
Connecticut State Department of Education
Room 350
Box 2219
Hartford, CT 06145
(203) 566-5401

DOCUMENT RESUME

ED 306 706 EA 021 046

AUTHOR Ralph, John; Dwyer, M. Christine
TITLE Making the Case: Evidence of Program Effectiveness in Schools and Classrooms. Criteria and Guidelines for the U.S. Department of Education's Program Effectiveness Panel.
INSTITUTION Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Nov 88
NOTE 57p.; The authors were assisted by contractor support provided by Research and Evaluation Associates and RMC Research.
PUB TYPE Reports - Descriptive (141) -- Guides - Non-Classroom Use (055) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Educational Change; *Educational Trends; Elementary Secondary Education; Evaluation Criteria; *Evaluation Methods; Government Publications; *Program Effectiveness; *Program Evaluation
IDENTIFIERS *Program Effectiveness Panel

ABSTRACT

Criteria and guidelines for the United States Department of Education's Program Effectiveness Panel (PEP), formerly the Joint Dissemination Review Panel (JDRP), are the focus of this report. The publication outlines procedural aspects of PEP's submission procedures and gives practical advice for projects seeking PEP approval. Chapter 2 answers questions about the submission and review process. Chapter 3 discusses changing trends in four areas: claims, case study methodology, types of evidence, and educational significance. Chapter 4 discussed the criteria of effectiveness applied by panelists. To complement PEP's formal criteria, Chapter 5 provides substantive guidance for panelists and program evaluators who want additional information about claim types and related evaluation concerns. The intent of Chapter 5 is to encourage both panelists and evaluators to reexamine their assumptions about evaluation designs and what constitutes convincing evidence. Chapter 6 describes the proper format for project submissions and gives advice about how to present information. Appended is a bibliography of 69 evaluation references. (SI)

EA

Making the Case

Evidence of Program Effectiveness in Schools and Classrooms

Criteria and Guidelines for the U.S. Department of Education's Program Effectiveness Panel

John Ralph
Office of Educational Research and Improvement
M. Christine Dwyer
RMC Research

November 1988

ACKNOWLEDGMENTS

A document of this nature has many antecedents and contributors because it is the product of ongoing discussion among those interested in PEP. This book has its obvious roots in *The Ideabook*, the original JDRP version of criteria and guidelines — but it is more than an update. *Making the Case: Evidence of Effectiveness in Schools and Classrooms* surveys and explains the many changes which have transformed JDRP into PEP. In the process of rethinking the panel's procedures, Ron Cartwright was an able critic and contributor.

Work on this version was begun under a contract with Research and Evaluation Associates. The authors were assisted by Sarah Roberts of RMC Research during that period. Work continued under the contract to RMC Research; review and assistance were provided by Trudy Brown, Nicholas Fitzgerald, Susan Klaiber and Gloria Zyskowski.

The final draft was carefully and thoughtfully read by five external reviewers who provided numerous suggestions for improvement which have been incorporated in the document. Those five reviewers were Judith Anderson, David Clark, Sam Corsi, Catherine Felknor and Joy Frechtling.

TABLE OF CONTENTS

CHAPTER 1

BACKGROUND AND PURPOSES OF THE PROGRAM EFFECTIVENESS PANEL

The Program Effectiveness Panel (PEP) is the Department of Education's primary mechanism for validating the effectiveness of educational programs developed by schools, universities, and other agencies. Based on evidence which applicants submit, the Program Effectiveness Panel (PEP) judges the difficulties of the goals which particular programs are designed to meet, whether those programs are effective in attaining their goals, and whether similar results are likely to be attained by others who use the program. Any project or product approved by the panel becomes a member of the National Diffusion Network and eligible to apply for federal dissemination funds, although other reviews are involved as well. Not every eligible project receives financial support for dissemination.

The credibility of PEP rests on two critical factors: its independence and its rigor. PEP is not authorized for the purpose of serving any particular division or organizational branch within the Department of Education. The Chair and all panel members are evaluation specialists approved by the Assistant Secretary for Educational Research and Improvement. A panel is asked to weigh the evidence of a program submission only after the program has been approved for review by one of the nine Assistant Secretaries in the Department of Education.

After receiving a program submission for review, the panel is charged with judging the program's effectiveness on the basis of evidence presented. Each program submission must formulate an empirical claim which attributes specific results to the program and presents objective and convincing evidence of the program's impact.

The forerunner of PEP, the Joint Dissemination Review Panel (JDRP), was created in 1972 as part of an effort to identify and make available programs which had been proven effective in local schools. In the beginning, only programs developed through federally funded program offices were eligible for JDRP validation. PEP review is no longer confined to programs developed with federal funds. Each year the panel reviews a broad range of programs from school districts and other agencies across the country which have received developmental funding from a variety of sources. Over the first 15 years, many changes gradually shaped and reshaped the panel's responsibilities and procedures. In 1987, many of these changes were formally recognized in the Final Regulations published in the Federal

Register notification 34 CFR 786, August 14, 1987. In brief, the changes (a) modified the panel's base of membership, (b) established new procedures for reviewing program submissions, (c) formalized criteria for assessing effectiveness, and (d) changed the panel's name to the Program Effectiveness Panel.

The membership of JDRP was originally an all-federal employee panel of approximately 25 members. Today, the Program Effectiveness Panel consists of approximately 60 members, all of whom bring expertise in program evaluation. About one-third of the members are employees of the U.S. Department of Education. The remaining two-thirds are drawn from universities and colleges, school districts, professional associations, and other entities concerned with educational research, evaluation, and practice.

Panel reviews of submissions are now conducted primarily by mail; if the need arises, in-person panels are convened in Washington, D.C. All panels consist of six members chosen randomly from the full membership. The panels are led by the Chair, who is an employee of the U.S. Department of Education and appointed by the Assistant Secretary for Educational Research and Improvement (OERI). PEP reviews are held as often as needed to review submissions received by the Department. About 50 submissions are reviewed each year; typically, panelists review projects in groups of three.

By regulation, PEP evaluates the strength of each program's claims of effectiveness by using specific categories. To determine a program's effectiveness, panelists assign points according to the merits and strengths of each program in the following areas:

- Results (0–50 points)
- Evaluation Design (0–40 points)
- Potential For Replication (0–10 points)

The points assigned to each area establish which factors are most relevant for the panel's approval or disapproval decisions. The results obtained by the program are singularly the most important criterion but results can only be convincing if the evaluation design is convincing. To be approved by PEP, a project's average total score must be at least 70 and it must receive an average of 40 points in the results category.

Now, as in 1972, the purpose of program review is to determine if particular programs work based solely on the evidence presented before the panel. PEP does not conduct program evaluations—it neither visits sites nor collects data on its own. Rather, PEP validates the evaluations done by others. An approval by PEP means that the evidence presented before the panel warrants the evaluation claims that a program achieves specific results. A disapproval does not necessarily reflect poorly on the program;

disapproval usually reflects poorly on the evaluation evidence. A disapproval means that the evidence presented to the panel does not warrant the evaluation's conclusions.

The remainder of this publication outlines procedural aspects of PEP's submission procedures and gives practical advice for projects seeking PEP approval. Chapter 2 answers questions about the submission and review process. Chapter 3 discusses changing trends in four areas: claims, case study methodology, types of evidence, and educational significance. Chapter 4 discussed the criteria of effectiveness applied by panelists. To complement PEP's formal criteria, Chapter 5 provides substantive guidance for panelists and program evaluators who want additional information about claim types and related evaluation concerns. The intent of Chapter 5 is to encourage both panelists and evaluators to reexamine their assumptions about evaluation designs and what constitutes convincing evidence. Chapter 6 describes the proper format for submissions and gives advice about how to present information.

CHAPTER 2

HOW THE PROGRAM EFFECTIVENESS PANEL WORKS

Individuals who seek review of educational programs or practices by the Program Effectiveness Panel (PEP) must compile the information required for a written submission which may not exceed 15 pages.

Program Office Review

Before a project's submission is reviewed by PEP, it must first be reviewed by a Department of Education program office. The federal program office that receives the submission conducts a preliminary assessment of (1) the project's evidence, (2) the degree to which it meets program office requirements, and (3) the submission's conformity to PEP guidelines. The office may also review the project for compliance with federal guidelines and accuracy of the evidence supporting claims of effectiveness. After its review, the federal program office recommends that the project proceed, that revisions be made in the submission, or that the project not be forwarded to the PEP. The federal program office which reviews the submission is determined by the project's original source of funding.

1. **If the program was funded through the Department of Education and the program office that provided funds is still in operation (such as Chapter 1, Chapter 2, or Special Education), the submission should be sent to that program office for review, approved by the appropriate Assistant Secretary, and transmitted to PEP.**

2. **Projects not funded by the Department of Education, or those projects for which the original program office no longer exists, should forward submissions to the Department of Education's Recognition Division (in the Office of Educational Research and Improvement).**

After receiving a project submission from the appropriate program office, a member of the PEP staff reviews the submission for completeness and schedules it for review by the panel.

All submissions are scheduled for a mail review which involves sending the submissions to six panel members for their comments and evaluation. Approximately 10 days later, a member of the PEP staff telephones each panelist to find out if the panelist has questions or needs additional project information before making a decision. When there are questions, the PEP staff member contacts the project developer, secures the requested information, and forwards the information to all panelists.

Voting Categories

As directed by the provisions contained in Rules and Regulations language of August 14, 1987, panelists evaluate submissions and award points on the basis of the following categories:

Results (0–50 points)

Panelists determine the extent to which the results indicate that:

- the program, product or practice's effect is convincing relative to similar programs; and
- the outcome claims of the program, product or practices are valid.

Evaluation Design (0–40 points)

Panelists determine the extent to which the evaluation design:

- is appropriate for the program, product or practice;
- is based on a correct interpretation of relevant research and literature;
- demonstrates that a clear and attributable connection exists between the evidence of an educational effect and the program treatment; and
- accounts for rival hypotheses that might explain effects.

Replication (0-10 points)

Panelists determine the extent to which the program, product or practice can be used at other sites with the likelihood of achieving similar results.

After the panelists complete their reviews, they each rate the submission in the above categories. The scores for each category are then added together for an overall rating between 0 and 100. The scores of six panelists are averaged for a final total rating.

Program effectiveness review approval is granted if the average panel rating for the *Results* category is at least 40 points, and the average total rating is at least 70 points. If the mail review results in a total average rating between 50 and 69 points, the Chair reviews the panel members' written comments to determine whether the vote represents a clear disapproval or whether further review by an in-person panel is warranted.

A second review (in-person panel) is justified if the panelists' written comments indicate a need for further clarity about the project's design or evaluation evidence. The composition of the in-person panel might not be identical to the original mail review panel. Typically, panels are convened in Washington, D.C., when there are at least three eligible programs for review. When an in-person panel is convened, the project developer is invited to attend the review to answer questions.

Mail reviews are usually completed within 6 to 8 weeks from PEP's receipt of the submission; project developers are notified in writing of panel decisions. A summary of the comments made by the panelists, including any concerns they had about the submission, is attached to the written notifications of both mail and in-person panel decisions. For further information about the PEP submission process contact:

U.S. Department of Education
Program Effectiveness Panel
Recognition Division
555 New Jersey Avenue, N.W., Room 510
Washington, D.C. 20208-5645
Telephone: (202) 357-6134

Additional materials about PEP, including dissemination process guidelines, are available through the above contact.

CHAPTER 3

HOW STANDARDS AND ASSUMPTIONS HAVE CHANGED

Through PEP, the Department of Education seeks to improve education by validating the effectiveness of specific educational programs and practices. To assure the Department and the public of the worth of these programs and practices, PEP must make balanced judgments which weigh the difficulty of achieving the program's goals against the strength of each program's evaluation design and its evidence.

There is a difference between an educational program that produces some effects and a program that is deemed effective. Education evaluators have well-established procedures for deciding if an educational program has an effect. Typically, measurement procedures for program effects are based on careful comparisons between changes in experimental and control group samples. The logic of hypothesis testing—determining whether an effect has occurred—only tells us if a program has made any difference at all. When we want to know a program's practical value for real-life classrooms—that is, when we want to know if the program is *effective*—we are asking different and more problematic questions about the importance of the effects observed. Decisions about effectiveness often require data from several sources and involve judgments that lie beyond the confines of statistical significance. For example, to weigh a program's educational impact, an evaluator may compare it with the results of similar programs.

Evolution of Standards of Effectiveness

The standards of effectiveness outlined in this chapter were shaped and refined by the operation and experiences of the forerunner of PEP, the JDRP. To understand current practices, it may help to place current standards in historical perspective.

During the 1960s, the infusion of federal funds into education sparked an explosion of innovations which resulted in many claims of effectiveness and much interest in disseminating the most successful programs. As Congress appropriated federal monies, the authorizing legislation frequently stipulated that formal evaluation evidence be gathered and reported back before reauthorization could occur. Some of the new programs proved themselves over time and became valuable additions to educational practice, while others turned out to be fads that wilted under the scrutiny of objective criteria and rigorous evaluations.

The Joint Dissemination Review Panel (JDRP), predecessor of PEP, was established to ensure that educational programs disseminated with federal funds had been properly evaluated and produced sound evidence of effectiveness. JDRP's function was quality control for program dissemination.

JDRP's insistence on valid and compelling evidence earned the panel's reputation for rigorous and exacting standards. Government education officials wanted to base decisions on the best and most reliable information even as they were faced with a bewildering variety of data and different evaluation standards. In that environment, JDRP held fast to classic experimental and quasi-experimental evaluation designs as a way to standardize the review process in light of vast differences in the quality of program evaluations. Over time, the panel's expectations became institutionalized, and extravagant claims supported by weak data—sometimes only by casual observation and testimonial support—were consistently rejected. Some misunderstandings, however, were more persistent, such as the idea that statistically significant effects alone were sufficient proof of the importance of educational innovations.

New Problems Emerge

As the field of program evaluation advanced, rigorous evaluation became an integral part of educational planning and development. In recent years, a new concern arose as a result of the emphasis placed on the experimental and quasi-experimental design approach. Developers worried that evaluation designs had become "the tail that wagged the dog;" the very standards that once brought rigor and consistency to JDRP's decisions had become inappropriate constraints for many innovative educational programs that sought the panel's approval. Critics charged that the elegance of the famous Campbell and Stanley approach to evaluation design was not tempered by the difficulties of measuring educational impact in working classrooms and schools. If, for example, a project sought to alter the disciplinary climate of a whole school, it was rarely possible for locally developed projects to collect data from an adequate control group (schools with a similar student body, comparable problems, and a willingness to participate) to meet the standards of experimentally designed evaluations.

In summary, for certain types of education programs, experimental and quasi-experimental evaluation designs imposed a narrow view of what constituted evidence. Some observers charged further that the panel favored programs with quantitative data and, more specifically, those programs with results that were measured by pencil-and-paper tests.

A review of the vast range of projects submitted over the years to JDRP supported the view that the panel needed a broader conception of appropriate evaluation designs and acceptable evidence. For example, in recent years, PEP reviewed a program whose goal was to involve students in highly acclaimed productions of Shakespearean theater. The developers of that program sought to instill something more than simple book

knowledge of Shakespeare's work—they wanted to engage the students in classical theatre and its rewards. PEP also reviewed a program designed to foster leadership skills in students and in another case reviewed a program designed to reduce athletic injuries in high school sports. In short, PEP reviews programs in many areas where schools want to excel and where schools have needs for effective strategies.

Current Assumptions

To encourage PEP's acceptance of a broader range of program claims and types of supporting evidence, several subtle but significant changes are incorporated into this revision of the criteria and guidelines. The first change is an understanding that not all projects must directly increase student achievement nor directly change student behavior; some programs may be effective at changing students' attitudes or teachers' attitudes towards a discipline (from which changes in learning should follow) or may aspire to alter the "academic climate" of a whole school. Second, these guidelines recognize that not all programs are ideally suited to an evaluation design which involves control groups, which relies on quantitative methodology, or both. Third, while PEP requires sufficient evidence to support the claims of a project submission, evidence can be of many different types. Finally, PEP has broadened its interpretation of educational significance, but its essential meaning to PEP has not changed.

Acceptable Claims

A review of changes in educational priorities, along with the variety of programs that routinely come before the panel, led to changes in PEP's assumptions about appropriate evaluative claims.

All programs reviewed by PEP are required to state claims. Those claim statements are expected to summarize the observable effects of each program. In the past, approvable claims were almost always limited to changes in student learning. This type of claim still typifies the kind of programs the PEP most often reviews, but new approaches have been added to encourage PEP panelists and program developers to think about claims which may focus on changes in attitudes and behaviors of teachers and students.

Instead of offering the traditional claims of increasing student learning, some programs may focus on objectives which are noteworthy because they are important advances toward difficult or long-term objectives related to learning. For example, an intermediate claim may focus on increasing students' interest in science courses or on increasing time spent on homework, either of which should lead to increased learning.

Other examples of nontraditional claims are those associated with drug abuse programs. A successful drug abuse program may, as a preliminary objective, undertake to change the attitudes of youngsters about drug use. As its ultimate goal, probably attainable over a significantly longer period of time and with additional resources, the successful program will aspire to reduce actual drug usage. Lowering the incidence of drug use would be the ultimate goal and final test of an effective drug abuse program. But for this

and similar programs, the attainment of an intermediate objective—changing the attitudes of youth about drug use—may be a difficult and momentous accomplishment.

In the case of the claim to alter students' learning, the essential requirement remains the same: Provide convincing evidence establishing the occurrence of change, as well as educational significance of the results. For the new types of claims related to intermediate objectives, the panel expects the evaluation design requirements and evidence to show that the intermediate objectives were achieved and are linked to important educational purposes. An example of an intermediate claim that does not meet PEP's standards would be the application of technology for its own sake—without evidence of a change in student attitude or behavior and without evidence of a convincing link to student learning as a result of the use of new technologies.

Qualitative Methodology: PEP and Case Studies

In PEP's history, only a few submissions have relied on qualitative methodology in the form of case studies for collecting evidence and drawing conclusions—even though there are circumstances in which qualitative approaches would be appropriate. When the effects of a project are complex, diverse, and subtle, a case study approach has strong advantages. For example, a project in which the treatment is tailored to varying circumstances at each site requires an evaluation which is similarly flexible. In the future, PEP expects some submissions will employ qualitative approaches to program evaluation, most likely in the form of case studies. For PEP's purposes, a case study can be defined as an evaluation based on comprehensive descriptions of complex situations, recounting what happened and why.

Case study methodology is an exciting approach for capturing educational program effects, but its very strength—the flexibility to portray programs accurately across diverse circumstances—makes the effort to meet PEP's standards a difficult challenge. Case studies, especially for the purpose of program evaluation, are costly, complex, and time-consuming because they require extraordinary efforts to pin down and verify a program's effects across several sites and across several complementary sources of data.

Case study methodology for the purposes of program evaluation must overcome the limitation of low generalizability due to small sample size. The strength of a case study is its familiarity with the particular workings of a program—knowing and relating how it is supposed to work, how it has worked, and the circumstances that explain past success or failure. Large-scale case studies which check for consistency across many sites become expensive and cumbersome. It is essential, though, for the panel to have evidence that a program can be adapted and work effectively in new sites. To this end a little imagination can sometimes go a long way. For example, evaluators may find that certain elements of the program are

most critical for its success and verify, through small-scale experimentation or unobtrusive observations, that similar programs with those critical elements are transportable and will indeed work as expected.

In reviewing case studies, PEP expects the submission to include: information about how sites were selected; identification of multiple sources and types of data; assurances that data collection has been comprehensive and unbiased; descriptions of analysis procedures that include attention to consistency in the evidence and tests for alternative interpretations. An evaluation's conclusions are more credible if one team of individuals collects the observational data and another team analyzes the field notes to draw conclusions.

Program developers considering the use of case studies for PEP submissions should read the overview written by Lois-ellin Datta (former chairperson of JDRP) of the General Accounting Office, *Case Study Evaluations* (April, 1987).

Variety of Evidence

In the past, the overwhelming majority of candidate programs relied on paper-and-pencil tests to document improvements in cognitive achievement. With the broadening of claim types, PEP expects to see greater variety in both evidence and methodological strategies for drawing conclusions from data.

No doubt most candidate projects will continue to employ standardized tests which (1) measure change along some predetermined scale and (2) have well-established reliability and validity. Even so, there are many forms of data—some less conventional—that are credible sources of evidence. For example, independent observations, questionnaires, official records (especially records that contain cumulative counts or logs of occurrences), and unobtrusive measures of many kinds allow for checks on reliability and validity and, yield information which often can be summarized by the usual statistical procedures. The scope of measurement and analytical procedures is limited only by the project's own constraints and the ingenuity of the evaluator. For all types of data, convincing the panel that the program evidence is compelling hinges on (1) successfully establishing the validity and reliability of instruments and procedures and (2) assuring the panel that collection and analysis techniques are credible.

PEP takes into account the special challenges posed by a particular educational setting or by an elusive goal. However, if evidence is thin or missing, simple excuses are unacceptable. It is not acceptable to offer scanty evidence due to the expenses of data collection, nor is it acceptable to plead lack of time and expertise. Similarly, the fact that no reliable instrument exists (to measure the program's outcome) cannot excuse the absence of sufficient evidence. Finally, the assertion that a program or practice is simply "important" cannot be substituted for objective evidence.

Educational Importance

Over the years, PEP has always been concerned about the critical distinction between program effects and the educational importance or significance of a program. Although the idea of educational significance may be interpreted more broadly now, its essential meaning has not changed. PEP is interested in both aspects of educational significance: effect size and substantive importance. Effect size deals with the magnitude of a given change, or how much of a difference it takes to make a difference. In education, one seldom expects no change in the absence of treatment. The phrase "no treatment expectation" refers instead to the amount of change that normally occurs as a result of a common treatment, maturation, or other typical causes. Effect size refers to the amount of change beyond that expectation. In quantitative studies, an effect size of roughly one-third of a standard deviation has often been accepted as the minimal practical difference.

When there is not an established "no-treatment expectation" with which to compare results, the second aspect of educational significance, substantive importance, necessarily plays a larger role. A program may be important because it addresses perennial school problems or contemporary social concerns, it demonstrates a high degree of cost efficiency, or it conforms to current government priorities.

The determination of educational importance must rely, in part, on normative judgments. Such judgments are made by the panel on the basis of data from similar programs which can be used to establish benchmarks. The determination of importance also relies on the relative effort required or costs incurred for implementation and maintenance. Regardless of the approach to establishing educational importance, PEP expects that all submittals will address importance as well as effects.

CHAPTER 4

CRITERIA OF EFFECTIVENESS

To assess the individual merits of each submission, PEP relies on general criteria for weighing educational effectiveness and specific guidelines for interpreting evidence. Both are necessary for PEP to assess the claims and supporting evidence for each program's submission. The general criteria for effectiveness indicate the kinds of questions all programs need to address; they lay the foundation for PEP's evaluative work. The specific guidelines give program evaluators and reviewers practical help in interpreting the general standards under varying conditions.

In order to be judged effective by PEP, all submissions must show that developers have met three general standards in the areas of evaluation design, results, and replication.

1. Evaluation Design

A credible evaluation design assures that the results have been obtained in a manner appropriate for the program and that the effects are clearly produced by the program.

Appropriate Measurement. An evaluation approach that meets PEP's standards of effectiveness relies on instruments and measurement procedures that are valid for the program and that have adequate technical strength. In effective projects, data collection and analysis procedures have been handled carefully; sufficient care is demonstrated for the reviewers to have confidence in the accuracy of results. Effective programs implement evaluation designs which are appropriate and reasonable even if only indirect measures of program impact are reported. The ineffective project usually errs by providing inadequate documentation about how and why measurement selections were made and about the appropriateness and strength of instruments and procedures. PEP attempts not to penalize projects from those fields in which the available instrumentation is limited or technically weak.

Attribution. Because PEP evaluates complex programs operating in real schools, clearly attributing results to the program is often the primary challenge for the program evaluator. In other words, it is critical for the program to select an evaluation approach that clearly demonstrates the link between program elements and observed outcomes.

The submittals judged ineffective by the PEP often fail to consider or convincingly rule out plausible alternative explanations for the observed results. An evaluation design which cannot test or control teacher effects, students' maturation, changes in related school policies, or selection

differences among program and comparison group participants is rarely convincing. Panel members expect program evaluators to know about potential threats to validity, to estimate the impact of competing influences when possible, and to recognize the design's shortcomings when alternative explanations cannot be ruled out.

Comparison Standard. An evaluation design should include an appropriate standard of comparison which clearly demonstrates the project's impact and the significance of that impact. In the typical case, PEP submissions compare carefully-drawn experimental groups which receive alternative treatments or they use norm-referenced test instruments to establish the effects of programs. Comparison standards are an essential design element for weighing the program's results.

2. Meaningful Results

The results of a program are meaningful when the impact is strong and the goals are important.

Programs often demonstrate value and importance by comparison to other programs or to alternative means of reaching the same results, but occasionally programs are considered effective simply because they have produced some results. For example, school programs that attempt to reduce juvenile delinquency or lower dropout rates may be considered successful when they show solid evidence of having made any inroads on these intractable problems. In these cases, the panel balances its judgment of effectiveness (based on comparisons to previous problem levels) against the difficulty of achieving the program's purposes. Frequently, programs demonstrate educational significance based on the program's efficiency, such as its ability to produce results in light of time, effort, and cost required.

To establish that a program has meaningful results—that is, valid and convincing evidence of useful results—evaluators should consider the need for the program results and comparisons with other similar programs.

Need. The PEP expects each submission to clearly state the need and purpose it fulfills. When interpreting results, the effective program makes an explicit connection between the changes observed and the practical needs met by the program. In some cases, a program's purpose may address a problem that concerns schools and districts everywhere; then, even a small practical effect may be important.

Programs which PEP judges to be ineffective often fail to consider outcomes in light of purposes, sometimes to the point of ignoring the obvious incongruity between stated goals and the measured effect. A surprising number of submissions that PEP reviews fail to adequately describe the program's basic purposes. It is both easy and trivial to demonstrate that students exposed to a particular curriculum will learn more about that

subject than those who are not. It is not enough to simply document a project's implementation and record its results. The panel must understand what need the program meets.

Worthwhile programs (for example, those featuring curricular enhancements) for which there is no pressing need can strengthen their arguments by demonstrating that students are not otherwise adversely affected—meaning that they lose nothing by their absence from other programs or activities when enrolled in the program. It is always necessary to provide an informed rationale for the overall value of the educational activity under evaluation.

Comparison to Similar Programs. All projects should provide accounts of how their programs operate and make clear distinctions among similar projects. An effective program is based on a clear conceptualization of what the program intends to achieve and how its particular approach succeeds better than other approaches. The program's design should reflect current research findings.

Conversely, submittals that PEP judges to be ineffective often make a simple and avoidable mistake: They fail to investigate how their programs compare to other programs of the same type. Such programs show little evidence of having learned from or built upon the efforts of others in the field. A failure to discuss the program's practical significance indicates to PEP an unfamiliarity with comparable programs in the same field.

It is a common misconception that every program that PEP considers must be innovative, that is, completely different from any other program. In fact, many projects that come before the panel are the result of an innovative approach which began in a local school. Innovation brings both advantages and disadvantages. From the PEP's perspective, innovations may be difficult to evaluate because there is no basis for comparison with similar programs. Also, innovative programs may pose methodological issues due to the confounding influences of local talents and prompt concerns about replicability.

3. Potential for Replication

The program must be transportable to other sites for reasonable costs—in dollars and effort—with the expectation of similar results.

PEP must determine if a program can be implemented at other sites for reasonable costs. It considers evidence of the program's generalizability and its efficiency.

Generalizability is usually measured by the stability of results at the home site or evidence of replication at new sites. An effective project demonstrates its generalizability by gathering comparable evidence across different settings or across several years. PEP is concerned about the context of all experimental sites so that it can determine where the program is likely to work—and work with some staying power. PEP expects

all submissions to identify the range of ages or grade levels, the populations, and the settings within which the program has been tested. The sample used in the evaluation should be adequate in terms of size and representativeness to support the claims.

Efficiency is measured by considering the money, time, and resources that the project requires—which include the demands made on both teachers and students—balanced against the program's results. A critical element of program efficiency is low or reasonable costs. The PEP must weigh a program's impact against the time, effort, or resources which the program requires. Detailed cost information for replication purposes and cost comparisons with competing programs are helpful to the panel. An effective program uses available resources efficiently relative to its results.

Realistic Expectations

No real-life program evaluations are wholly convincing, and rarely are they totally unconvincing. The average submission meets several requirements of the general criteria quite easily and has difficulty addressing the others. The best advice is to remedy as many design problems as possible, often by collecting supporting evidence to complement the basic evaluation design, and to be frank and thoughtful about remaining shortcomings and uncertainties. Ultimately, what makes program evidence convincing depends on (1) the difficulties of achieving the program's goals and (2) the difficulties of measuring the program's results. PEP's expectations are crafted to be both realistic and rigorous. In every case, the panel expects program developers to be aware of the problems—solvable and unsolvable—in their evaluation designs and to demonstrate that every reasonable effort was made to obtain compelling evidence of the program's effectiveness.

CHAPTER 5

TYPES OF CLAIMS

The expansion in the types of claims and the nature of evidence appropriate for consideration by PEP has led to the development of four claim types to guide program evaluators. For each type, examples and discussion provide guidance for design of evaluations and presentation of supporting evidence. PEP's standards for reviewing each claim type are included in the form of questions.

In practice, a submittal may include claims that are from two or more types suggested here or that represent combinations of claims. Typically, projects focus on one type of claim and provide supporting evidence that may relate to another claim. For example, a project may claim student achievement change and then supplement its primary evidence with indications of related student attitude changes.

Claim Type 1: Academic Achievement—Changes in Knowledge and Skills. This is the traditional claim, usually based on experimental or quasi-experimental evaluation designs. It requires measurement of learning and the comparison of growth to an appropriate control group or normative standards. It also requires a convincing demonstration that overall change is educationally significant.

Claim Type 2: Improvements in Teachers' Attitudes and Behaviors. Claims of this type focus on programs that change teachers' attitudes and behaviors in order to improve the teaching process. They require demonstration of changes in attitudes or behaviors, and presentation of a reasonable link between these results and an educationally important goal.

Claim Type 3: Improvements in Students' Attitudes and Behaviors. This claim type focuses on changes in students' attitudes and behaviors that in the long term lead to educationally desirable outcomes. Use of this claim requires data showing positive change in the target group, and strong logical or empirical evidence that this change is large enough to be educationally meaningful.

Claim Type 4: Improvements in Instructional Practices and Procedures. Claims of this type are intermediate outcomes that have to do with such system changes as efficiency, cost and labor savings, and improved services. This claim type requires documentation of change and demonstration of the link to longer-term educationally relevant outcomes.

Additional claim types will be developed as other types of programs seek PEP approval.

Claim Type 1 — Academic Achievement: Changes in Knowledge and Skills

Traditionally, programs claiming to result in greater knowledge or increased learning of skills have been the most likely to come before the panel. The claim may demonstrate gains in knowledge or skills by any type of learner—students at any grade level, teachers, or other adult learners.

Projects for which this model is most appropriate are instructional interventions that teach content or skills, or provide opportunities for students to apply knowledge. Examples are traditional school curriculum areas such as reading or mathematics and emerging subjects such as computer science and thinking skills, as well as areas such as adult literacy. Claims in this area are based on the observation of measurable changes in the target population.

Examples of Claims

- Acquisition of factual knowledge: Students in the physics project at three typical high schools made greater gains than the national norm group on a standardized test of physics knowledge.
- Acquisition of new types of knowledge (*i.e.*, knowledge not presented in a typical curriculum): When compared with a control group, students in a computer literacy course scored significantly better on reliable (split-half r = .93), locally developed tests of computer knowledge.
- Rapid acquisition of knowledge (*i.e.*, changes in the efficiency of learning): Students completing a 1-semester math course performed as well on a standardized test as did a matched comparison group of students taking the traditional 1-year course.
- Application of knowledge: In addition to making greater-than-expected gains in library reference skills on a standardized test, program students required significantly less assistance with research activities than comparison students, as measured by structured observations in school libraries.
- Acquisition of skills: Quantitative studies in eight separate sites, using various nationally known measures, showed significant advantages in the area of reasoning ability for students in a philosophy program over comparison group students.
- Application of skills: Project students achieved significantly better ratings on analytically scored writing samples than did comparison students in the regular language arts program.

Projects offering this type of claim often present evidence based on familiar measures. Chief among these are written tests, including standardized norm-referenced tests, locally developed tests, and criterion-referenced tests. Generally speaking, tests have the advantage that their reliability and validity can be determined using established psychometric methods.

Publishers of major standardized tests provide this type of information for their national norming samples and for various subsamples, which is one reason why such tests are the most commonly used instruments.

Similar to objective tests are direct ratings of performance such as holistic or analytic ratings of writing samples. Reliability of ratings can be determined using established methods. Other measures of performance include structured interviews or observations of skill demonstrations and content analyses of student work. Compared to written tests, these measures require greater effort in administration, scoring, and interpretation, as well as in training for those who administer them. This is a major reason why they are less commonly used; however, with methodical and appropriate implementation, an evaluation using such instruments can be more convincing than one using nationally standardized tests.

The classic experimental or quasi-experimental design is most frequently used for this model. Its essential feature is the comparison between performance of the program treatment group and an appropriate comparison group. The easiest variation of this design is one using national norms; however, this comparison is appropriate only when the treatment group is similar to the national norm group on educationally relevant variables. Since target groups are often restricted in terms of one or more variables (for example, income, achievement level, race or ethnicity), it may not be possible to meet this requirement. For this reason, a better matched local comparison group is often used.

Evaluation Design

The primary concern of the panel in a Claim Type 1 evaluation design is the comparison of the treatment group's performance to some appropriate "no-treatment" expectation. Although designs using a comparison group or norm group are the most common, there are also some designs, such as time-series designs or multiple-baseline designs, for which the expectation can be derived from pre-project growth rates established for the subjects themselves. The appropriateness of the comparison is the critical factor. If the selected comparison group is *not* similar to the treatment group in one or more relevant ways, or if there is some reason why performance of the subjects in a multiple-baseline design may not be linear, then the expectation may not be valid.

Potential Panel Questions About Evaluation Design:

- Is this the strongest and most appropriate research design that could be undertaken given the nature of the project treatment, setting, and participants? If not, what are the reasons for not choosing another design?
- Have the inherent assumptions of the design been taken into consideration?
- Can the appropriateness of the comparison standard be demonstrated?
- How was the comparison group chosen? Is there evidence that it is similar to the project group in educationally relevant ways?
- If participants were selected on the basis of test scores, has a separate pretest been used in order to avoid the regression effect error? Have other measures been taken to counter the impact of regression?
- If a sample of program participants is used, is it a representative sample and has the sample been selected in a nonbiased fashion? Is the same true for the comparison group?
- Is the size of the evaluation sample large enough to generalize with sufficient confidence to the target population as a whole?
- Have sufficient numbers of learners remained in the study during the treatment period? Have the reasons for attrition and its effects been investigated?
- Have participants been selected in accordance with rules for the evaluation design?
- Is the timing of data collection appropriate and logical for the treatment and for the instruments used?

Instruments, Procedures, and Data Collection

The actual methods used to measure the changes produced by the project are of great importance, since no amount of analysis or argument can redeem a body of evidence that is flawed by improper choice of instruments, incorrect procedures, or contaminated data.

The panel needs sufficient information showing that the instruments are appropriate, reliable, and valid as measures of the project's claims. The less well-known the instrument, the greater the burden for the evaluator to establish these points. Whatever the instruments, it is important to indicate that they were administered in the proper way for both treatment and comparison groups. Possible sources of contamination in the data should be guarded against; if they are unavoidable, their effects should be acknowledged and, if possible, estimated.

Again, the less straightforward the data collection procedures, the greater the burden for the evaluator to document that procedures are credible. For example, if writing samples are used to demonstrate improvement in composition skills, they should be typed so that raters will not be influenced by penmanship or extraneous appearance factors, and pre- and post-writing samples should not be identified as such to avoid creating differential expectations of the two samples. All scoring should be done after the post-data is collected. Pre- and post-writing samples should be

randomly intermixed so that any rating effects are distributed across both sets of data. If possible, independent raters should be used rather than persons familiar with the writing of either project or comparison group students.

Potential Panel Questions About Instruments, Procedures, and Data Collection:

- Is there evidence that the instruments or procedures are valid for the treatment?
- Is there evidence that the instruments or procedures are reliable?
- How directly do the measures relate to the submission's claims?
- Are the measures accepted in the field? If they are not well known, why were they selected or how were they developed?
- Were test levels appropriate for participants? For example, have test floor and ceiling effects been avoided?
- Have steps been taken to control for practice effects? For example, have alternate forms been used?
- Were instruments administered under standardized conditions?
- If norm-referenced tests were used, were tests administered during correct times?
- If tests are not objectively scored, were steps taken to ensure impartial and reliable scoring?
- Were correct procedures used for converting raw data to derived scores?

Analysis and Discussion of Results

If the research design and data collection aspects of the study have been handled properly, the results emerge clearly. The panel is looking for correctness, clarity and plausibility of results. The statistical analyses selected must be appropriate for the type of data and the number of cases. Analysis procedures must correspond to the rules of the evaluation design and should take account of any special circumstances that may have occurred. For example, if treatment and comparison groups differed significantly on their pretest means, appropriate correction procedures should be used in the analysis and in the presentation of results.

Results should be presented in a manner that is clear and appropriate for the data. Care should be taken not to over-aggregate, for example, by combining gains across all grade levels in primary basic skills projects. Such a combination is likely to raise the suspicion that the average gain reflects outstanding performance by one grade level while masking losses at another. For clarity, standard deviations should be included. If space allows, it is desirable to present enough data to allow for independent checks of figures. If unusual formulas are used, they should be provided. Clarity is enhanced by careful presentation of data summaries, using tables, graphs, or charts which should be legible, clearly labeled, and complete but not overcrowded.

In the case of large-scale projects presenting data from a number of separate evaluation studies at different sites, a somewhat different approach to the presentation of data is required. Special care should be taken that each evaluation study included in the submission meets all of the criteria already described for suitability of design and implementation. Salient features such as type of design, description of subjects, measures used, and duration of the study, as well as the nature and significance of results should be summarized for each site in narrative or chart form.

The presentation and interpretation of results should be plausible. Because cognitive achievement is one of the oldest and most established areas of measurement and evaluation, and of panel review, there is a reasonable expectation about the growth that can actually be achieved. A basic skills project that claimed an effect size of one and a half standard deviations should re-examine its entire evaluation process for possible errors. Even when the size of gains is plausible, evaluators should be wary of possible alternative causes of observed effects. Where appropriate, the submittal should address rival hypotheses. Finally, in summarizing project results and drawing conclusions, the evaluator should bring together the major points that support the plausibility of the linkage between the project treatment and ultimate educational importance.

Potential Panel Questions about Analysis and Discussion of Results:

- Did analysis procedures fulfill the requirements of the evaluation design?
- Were analysis procedures appropriately chosen and properly carried out? For example, were appropriate statistical tests used?
- Were appropriate scores used?
- Are the results consistent across observations?
- Do the results show statistical significance?
- Are the effect sizes large enough to have practical significance? Are they plausible?
- How does the size of gains compare with those from comparable treatments on similar populations?
- Are there negative or positive effects on learning in subject areas unrelated to the submission claims? For example, was less time spent on other subjects to allow intensive treatment in the project area?
- Are the observed effects accounted for by rival hypotheses, such as:
  - differences in teacher ability, experience, or charisma?
  - treatments outside the program that affected the program group?
  - teaching to the test?
  - Hawthorne effects?
  - maturation?

Claim Type 2: Improvements in Teachers' Attitudes and Behaviors

Many projects seek to improve teaching and learning by influencing teachers' attitudes and changing their teaching behaviors, or both. Claim Type 2 should be used if the project meets both of the following conditions:

- It is aimed at the intermediate effect of producing changes in the attitudes and behaviors of teachers; and
- It postulates that these changes will contribute to student achievement some time in the future.

Typically, claims of this type are intended to achieve these changes in a targeted participant group rather than in an entire institutional population.

Claims are based on observation of measurable change in the targeted participant group.

Examples of Claims

- Increase in the amount of instruction devoted to a subject: After implementation of the new hands-on science program, participating teachers reported an increase of at least 20 minutes per week in the time devoted to science instruction, while non-program teachers showed no increase ($p<.01$). Pre- and post-classroom observation figures confirmed this finding.
- Increase in total instructional time: Teachers who participated in the computer management project reduced the time spent on recording attendance, tardiness, homework completion, lesson assignment, and test scores. Classroom observations showed that they increased their time spent on direct instructional contact by one-third standard deviation over a 1-year period, and maintained this gain throughout the following year.
- Change in instructional methods: This program produced changes in teachers' instructional strategies for teaching Shakespeare, including greater interest and enthusiasm for the subject, and greater use of methods emphasizing student participation in actual dramatic performance. These effects were documented through questionnaire response and through voluntary teacher participation in the program, which over 7 years increased from 30 to 100 teachers at the elementary level and from 30 to 150 teachers at the secondary level.
- Change in emphasis within a discipline: Social studies teachers who participated in the research and problem-solving workshops modeled problem-solving approaches more frequently in the classroom and gave more assignments requiring use of research skills than did a comparison group of teachers drawn from the same schools. After 1-year, teacher questionnaires, student questionnaires, and pre- and post-classroom observations all showed statistically significant differences.

- Positive change in teachers' expectations of students: After attending a summer institute on "How to Improve Student Writing," high school English teachers exhibited higher standards for student compositions. There was a statistically significant increase in the number of composition assignments. Using random samples of completed student assignments taken at uniform intervals over the pre-/post-program year, evaluators also found a statistically significant increase of one-half standard deviation in the number of mechanical errors students were required to correct, and the number of teacher comments on content, organization, and style.
- Positive change in teachers' expectations: The project increased teachers' expectations of elementary students' involvement in classroom activities. Structured pre/post observations showed an average two-fifths standard deviation increase for project teachers over comparison teachers in the number of different students called upon during a lesson, use of techniques to check student attentiveness, and use of techniques for bringing "wandering" students back on task. The observations also revealed fewer discipline problems for project teachers. These differences were maintained for a 1-year period.

Measures employed for claims of these types are likely to include (1) response-soliciting instruments such as interviews, surveys, and questionnaires, (2) data from structured observations, or (3) unobtrusive measures such as counts and statistics of all kinds, or data from school or district historical records. Questionnaires are the most frequently used. Their forms are many and varied, including multiple-choice items, rating scales, and free-response questions. One strength of questionnaires and structured interviews is that they can be designed to address specifically the question under study. Rather than hypothesizing a relationship between observed behavior and a particular knowledge, attitude, or opinion, the evaluator can ask the question directly. The drawback is that the answers are not necessarily unbiased, but may be colored by the respondent's perception of the evaluator's purpose or other extraneous influences. Therefore, these are termed "reactive measures."

As reactive measures, the potential for respondent bias is so great that questionnaires should not be relied upon as the sole source of support for a project's claims. A successful submittal will typically present data from other sources in preference to, or in addition to, questionnaire data. Sources include structured observation data and data collected unobtrusively over an extended period of time (such as records in the form of teaching plans and assignments).

When response-soliciting instruments are used, evaluators should recognize the potential for error that lurks in an all-too-often neglected area—nonresponse bias. That is, it is difficult if not impossible to generalize results when only a small portion of the targeted group actually returns questionnaires. Professional pollsters and survey researchers have refined their techniques of sampling, item design, incentives, and followups

to the point where the range of error in their results is typically between 1 and 5 percent. They can make this claim because they have either ensured a high rate of return, or they have used sophisticated methods to estimate nonresponse bias. Studies done in schools, however, sometimes have a response rate as low as 40 percent. In such a case, even a claim that 90 percent of respondents express satisfaction with a program could still mean that up to 64 percent of the population might disapprove of it.

Structured observations offer a way to assess changes without relying on self-reports, although they may be influenced by the subject's awareness of being observed or by observer biases. These problems can be avoided to a large extent by using several observations over a period of time and by careful training and field-testing to ensure reliability of the instrument and the observers.

Designs for this claim type include experimental or quasi-experimental designs using either an appropriate comparison group of teachers who did not receive the treatment, or data on the attitudes and behaviors of teachers before and after receiving the treatment. In either case, evidence is usually collected over a long enough period of time, typically a year or more, to indicate that the change is sustained and stable. To strengthen the submittal, projects have presented evidence of growth of demand for the treatment over a number of years, or of its spread to locations beyond the original site.

Evaluation Design

The criteria for sound evaluation design described for the previous claim type also apply here. As always, the provision of an appropriate no-treatment expectation is the key. There are, however, some distinguishing points with this model that evaluators should bear in mind. Claim Type 1, which deals with claims of direct change in cognitive achievement, has the advantage of a relatively long history of evaluation research with a well-established consensus about what constitutes important change. With intermediate claims of specific teacher training or support, however, there is a need for the evaluator to pay special attention to the validity of the no-treatment expectation used in the design. Evaluators should ask questions of the following type. How much teacher "response" (as measured by expressions of enthusiasm, initial use of methods, or growth in enrollment) should be expected of any program simply because it is new and different? Should we expect a certain "newness effect?" Have the developers created an important and durable contribution to teaching, or simply another fad? To some extent, a repeated measures or large-scale design can help to offset this problem. In addition, the question of magnitude of change should be addressed with logic and common sense in relating project outcomes to educationally important goals.

Potential Panel Questions About Design:

In addition to questions raised regarding sound design principles under Claim Type 1, the following apply:

- If comparison was between groups of teachers, how was their pre-treatment equivalence documented?
- If the evaluation used a sample of teacher participants, how was the sample selected? How representative is the sample of the participant group? How valid is the comparison sample?
- To what extent are the teachers representative of the general teacher population in terms of background, training, and experience?
- Are the samples large enough to generalize with confidence to the population as a whole?
- Is the sample large enough to have confidence in the reliability of the observed effect?
- Are selection methods unbiased, as opposed to having the treatment group formed of teachers who volunteered and the comparison group of those who did not volunteer?
- What was the response rate of self-reporting measures? How was nonresponse bias addressed?

Instruments, Procedures, and Data Collection

Evaluations of these claims are usually focused on an identified group of program participants rather than an entire population. This is likely to mean use of measures specially administered for the study, although extraction and use of appropriate institutional records is also a possibility.

In assembling the battery of measures to be used, the evaluator should carefully consider a number of common sense questions: If the hypothesized change is taking place in teachers, how would we see it? How many different kinds of indicators can we identify? Which of these can be measured by unobtrusive means (the most objective)? Which by systematic observation (some possible subject or observer bias)? Which by self-reporting instruments (the most subjective)? After making a list of possibilities under each of these categories, appropriate choices can be made, taking into account the combinations of indicators which would be most scientifically sound, most persuasive to an outside review panel, and most feasible to implement given the resources available for evaluation.

Evaluators should not only present information regarding the appropriateness, reliability, and validity of the instruments, but also document the adequacy of data collection procedures. For example, if observations or interviews are used, the training of data collectors and the means used to determine their reliability should be described.

Potential Panel Questions About Instruments, Procedures, and Data Collection:

In addition to questions raised for the previous claim type, the following apply:

- How well does the attitude or behavior measured by the instruments correspond to the underlying treatment construct?
- How reliable are the instruments and data collection procedures?
- If self-report, what cautions have been taken to ensure objectivity?
- If observations are used, how is the observation schedule related to implementation of the treatment? Have multiple observations been used to measure stability of results?
- Has there been attrition from the sample(s)? What are the reasons for attrition? What effect might this have had on results?
- If the major measures rely on self-report, what other evidence corroborates this data?

Analysis and Discussion of Results

General criteria are the same as for the previous claim type. For intermediate claims, three points deserve special emphasis: effect size, cost, and reduction of other opportunities. It may be tempting to imagine that because achievement is not being addressed, the demonstration of any positive change is adequate, and that the importance of that change need not be established. In fact, these claims are subject to the same demand that is placed on achievement claims—that results are of sufficient magnitude to be both statistically and educationally significant. If the difference between the observed and the expected can be measured, the size of that difference can be expressed in the familiar terms of standard deviation units. Further, the size of change should have a common sense value. A difference between program and comparison teachers in the amount of time devoted to a subject might be statistically significant, but if the difference amounted to only 5 minutes a day, would it be likely to have educational importance?

A second point to consider is program cost. When the relationship between program outcomes and ultimate educational goals is based on a strong theory rather than empirical evidence, the cost of achieving the outcomes can become a larger factor in assessing program importance. Evaluators should point out any advantages a program may have in terms of actual dollar costs, as well as time required for planning and implementation.

Third, opportunity costs (*i.e.*, foregone opportunities) must be considered. Claims often carry a clear implication of opportunity costs, and these should be addressed. Consider claims of "increase in amount of instruction devoted to a subject" or "change in emphasis within a discipline." From what other subjects is the time being taken to devote to the program subject? Is emphasis on something important reduced in order to emphasize something new?

In addition to these cautionary points, evaluators should consider the general likelihood of demand for the program, how it fits in with current educational priorities, and its potential for replication. Since programs are aimed at teachers, it is appropriate to examine the applicability of a particular change to the teaching process as a whole. For example, how many teachers, subject areas, grade levels, or localities could profitably be influenced by the program?

Finally, as for all programs making intermediate claims, special attention should be paid to the hypothesized link between the immediate program outcomes and student achievement. There is a critical distinction between the program effects and the program's educational importance. The existence and size of program effects should be established by measurement. This is the *raison d'être* of the experimental design model as well as the associated data collection and statistical analyses. In the case of intermediate outcomes, educational importance must usually be supported by logical argument, experience or tradition, common sense, or expert consensus. It is important not to confuse effects and importance by using the wrong method in the wrong place; for example, by relying on testimonials to do both jobs.

Potential Panel Questions About Analysis And Discussion Of Results:

In addition to questions raised under previous models regarding correctness, clarity, and plausibility, the following apply:

- What is the range of situations in which the results have been observed? Different departments? Different school organizations? Different types of communities?
- Are there any unintended benefits from the program? For example, are there positive carryovers into other teaching areas?
- Are there any unintended negative effects from the program? For example, have time or resources been taken away from other disciplines?
- Are the observed effects accounted for by rival hypotheses, such as:
  - other training, incentives, or requirements that affect teachers?
  - unique or unusual characteristics of the project school(s)?
  - Hawthorne, "status symbol," or "halo" effects from participation in the project?
- What is the hypothesized link between the changes in teacher attitudes or behaviors and the ultimate impact on student learning?
- How long-lasting are the observed changes?

Claim Type 3: Improvements in Students' Attitudes and Behaviors

Projects with goals of improving learning sometimes claim to change students' attitudes or behaviors as a foundation for future achievement. If the project's goals and claims are focused on the ultimate learning outcomes (*i.e.*, increased academic achievement) resulting from improved student attitudes and behaviors, then Claim Type 1 is the appropriate evaluation strategy. Claim Type 3 should be used if the project meets both of the following conditions:

- It is aimed at the immediate effect of producing changes in the attitudes and behaviors of students; and
- It postulates that the outcomes will contribute to student achievement some time in the future.

Typically, claims of this type are intended to achieve changes in a targeted subgroup rather than in an entire institutional population. Claims are based on observation of measurable change in the targeted group.

Examples of Claims

- Increase in attendance: Students in the program showed an attendance rate significantly higher than that of the comparison group during the program year, and this gain was maintained during the year following treatment.
- Decrease in drop-out rates: One year after implementation of the program in all classrooms, the school dropout rate had fallen by 0.4 standard deviation units. No change was observed in similar schools nearby.
- Positive attitude about learning, school, self as a learner: With increased use of the program's learning kits, students demonstrated significant increases in ability to identify the purpose of their learning activities, awareness and use of a variety of resources and materials, application of information in project-related activities, and enthusiasm for and involvement in library media center activities, as indicated by structured interviews.
- Change in attitude about the value of a subject: After participating for 1 year in this science program, students exhibited more favorable attitudes toward the learning of science than students in the traditional program ($p<.01$), as measured by a questionnaire. Parent and teacher questionnaires supported this result. The number of students seeking additional work in science outside of school increased.
- More realistic course selection for career direction: Student choice of elective courses matched vocational preference profiles better than before the program, based on independent judgments of three raters using a contingency table showing the relationship between courses and job categories.

- Rise in academic level of courses students select: During the program year, participants in the study counseling project were more likely ($p<.01$) than comparison students to enroll in courses designated "academic." They were also more likely to complete the course, and received better grades (chi square <.05, df=4). This trend was maintained over 2 subsequent years after completion of the program.
- Improvement in course completion rates: This program, aimed at students who had previously dropped out of high school math courses, resulted in a 0.3 standard deviation increase in the number successfully completing the basic math requirement.
- Change in nutrition or health-related behavior (*e.g.*, smoking or drug use): One year after the program began in all classrooms, school records indicated a decline in the monthly average of incidents of drug possession, drug use, or drug selling on campus.
- Physical change (*e.g.*, weight loss or fitness): Based on performance measures adapted from the military, students in the program improved their fitness scores significantly more than students in the comparison group.

Like Claim Type 2, this model typically does not directly involve academic achievement. To substantiate results, programs may assemble a body of corroborating evidence from various sources. Thus, several different types of instruments are often used in combination, including questionnaires or other self-reports, observations or interviews, and unobtrusive measures. Examples include school records; case studies of individual students, classes, or schools; structured interviews with students, parents, teachers, community service agencies, or police; and post-program followup of students' college or job choices.

Designs are similar to those used for Claim Type 2. Pre/post measurement is typically used. The no-treatment expectation is derived from a similar comparison group that does not receive the treatment, from longitudinal data on the attitudes and behaviors of students before receiving the treatment, or by examining pre/post data for different grade or age levels to estimate maturation effect.

Evaluation Design

The basic criteria for sound evaluation design are the same as in previous models. A valid no-treatment expectation is the critical factor, and, as with intermediate models, a successful submittal must present complete information, including some that might be taken for granted in other models. For Claim Type 1, the national norm sample of a major test publisher represents the national student population. If a Claim Type 3 design uses "national averages" however, it is important to specify the source of these figures, how current they are, and what kind of group they represent. For example, it would not be appropriate for an antidrug program in a small suburban high school to compare results with a national average that was based on large urban high schools.

If possible, the research design should provide for multiple measures—which do not share the same biases—of intended program outcomes in order to control for possible biases and build a convincing body of evidence. If claims are based on evidence from self-reporting instruments, such as questionnaires, it is particularly important to use additional measures that can provide corroborating evidence. Multiple comparisons can be useful, provided each one is appropriate. For example, results can be compared with those for a comparison group and also with the pre-program trend for the target group.

Potential Panel Questions About Design:

In addition to questions raised for previous models, the following apply:

- Has the design used an adequate method of estimating what would have happened without the treatment (no-treatment expectation)?
- If a comparison group was used, how was similarity to the program group documented? Were pre-treatment data obtained for both?
- Is a variety of measures (including variety in source of information) and comparisons used?
- Are measurements taken at critical points? Are measurements taken often enough to estimate stability of results?
- If sampling was used, how representative are the samples of the program and comparison groups?

Instruments, Procedures, and Data Collection

Project-specific measures will usually be required. In designing such measures, every attempt should be made to ensure objectivity. Student or teacher questionnaires that essentially solicit testimonials for the program should be avoided. For example, an item that asks respondents to list the three most helpful things about the program may be useful for formative evaluation purposes, but it is not adequate as evidence of effectiveness. For some program areas, there are recognized "ready-made" measures that may be considered; for example, scales measuring attitudes toward various school subjects, vocational preferences, and self-concept. As always, reliability and validity are important.

In assembling a convincing body of evidence, use of multiple measures can be an asset. Different kinds of instruments can be used to assess a particular kind of change. For example, a change in student attitudes about the value of a subject could be measured by questionnaires, by observations of classroom participation, or by unobtrusive measures such as number of students enrolled, or number of students pursuing related activities outside of school.

In addition to using different kinds of instruments, a single instrument can be used to get information from different sources. For example, a questionnaire could be administered to students, to teachers, and to parents. The responses of each group could be used as separate measures

of change. Another option would be to use responses from a parent or teacher cross-validation sample to increase confidence in the reliability of responses from the student questionnaire.

Potential Panel Questions About Instruments, Procedures, and Data Collection:

In addition to questions raised under previous models, the following apply:

- Is the sample of the attitudes or behaviors measured representative of the outcome claimed? How is this validity demonstrated?
- How reliable are the measures used?
- If interviews or observations were used, what was done to ensure objectivity of the interviewers or observers?
- How were observers trained to ensure that the same attribute was seen?
- Were instruments administered under standardized conditions?
- If measures are self-reports, what steps were taken to promote objectivity?
- What kind of response rate was obtained? How was nonresponse bias addressed?
- If ratings were subjectively assigned (*e.g.*, grades, placements), was care taken to ensure systematic application?

Analysis and Discussion of Results

As with other intermediate claims, it is important to remember the distinction between two components of educational significance: (1) effect size, which is to be measured as scientifically as possible through sound experimental design and appropriate statistical analysis, and (2) substantive importance, which depends heavily upon judgments of value. Clearly, programs with intermediate goals have a bigger job to do in presenting a compelling case for substantive importance, compared with programs aimed at academic achievement. This does not mean that a submittal can afford to downplay the presentation and discussion of actual data. No amount of fashionable aura around the program concept nor glowing testimonials can substitute for an objective demonstration of real change. As with Claim Type 2, actual costs and possible reduction or limitations of other activities should be addressed.

Potential Panel Questions About Analysis and Discussion of Results:

- Are the samples representative? Are they large enough to generalize with confidence to the population as a whole?
- Are results differentiated by student characteristics such as age, sex, ethnic identity, ability?
- Is there a plausible relationship between the nature of the treatment and the effects claimed?

- Is there evidence that the effects are sustained, or do they diminish after completion of the treatment?
- Has the program examined unintended outcomes as well as intended results?
- What is the hypothesized link between what is measured and student achievement?
- Do experts agree that the change suggested by outcomes is an important one?
- Are there harmful effects?
- Are there rival hypotheses that could account for the observed change in student attitudes or behaviors? For example, might the change be attributed to:

  - other programs, sanctions, or incentives?
  - outside social influences or larger societal trends?
  - specific local events, such as student deaths related to behaviors later targeted by the program?

Claim Type 4: Improvements in Instructional Practices and Procedures

There is another class of projects in which the goals have to do with changes in the education system—its efficiency, the types of services it provides, or coordination among its different elements. Examples might be programs that reduce costs, save labor, promote interdepartmental cooperation, provide new types of services, or improve services to particular client groups. Such projects may operate in schools or in other institutions with education-related missions and links to schools, such as libraries and museums.

If the goals and claims relate to intermediate changes in a specified participant group (*i.e.*, teachers or students), then Claim Types 2 or 3 should be used. Claim Type 4 is appropriate when the project meets the following conditions:

- It is aimed at the immediate effect of producing changes in the school, system, or institution, and/or changes in a general population or service area;
- It consists of a coherent set of procedures that can be transferred to similar institutions; and
- It postulates that the outcomes will contribute to student achievement some time in the future.

PEP is interested only in those claims that are related to student learning (either directly or in intermediate fashion) and not simply in improved functioning of educational institutions. For example, PEP understands that efficiency in plant maintenance or cafeteria food savings are important objectives for schools but such types of changes are outside the scope of PEP's mission.

A Claim Type 4 evaluation seeks to demonstrate the achievement of immediate goals that produce a change in a given system's delivery of academic services or in the target population's use of services, or both.

Examples of Claims

- Improvements in service to particular client groups: Through the satellite video science service, children at six hundred sites participated in the master teacher science lesson series for a cost of approximately $1.00 per student.
- Reduction in costs and improvements in efficiency of service delivery: By instituting a cooperative program among three school districts, duplicate arts education efforts were eliminated and expanded opportunities were provided at significantly reduced costs.
- Increase in use of information: As a result of the program, average monthly circulation of science-related kit materials to teachers increased by one-half standard deviation over monthly figures for the pre-program year.

- Increase in use of resources and facilities: One year after conversion of a neglected branch library into a "homework library" staffed by teacher-librarians and stocked with young-adult level materials, monthly figures for library visits quadrupled, the number of library cards issued doubled, and circulation was three times larger.

Evidence of improvements in practices and procedures will often be found in the form of existing records. With careful planning, special recordkeeping procedures may be instituted as a program begins in order to measure change. Types of records include:

- projected and actual budgets
- records of expenditures
- records of staff utilization
- participation, enrollment, and attendance counts
- materials circulation records
- number of requests received
- number of requests filled
- response time records

Questionnaires, surveys, interviews and structured observations are sometimes used when new or additional data must be collected. Evaluation designs usually rely upon pre/post measurement. A no-treatment expectation is derived from the previously existing conditions, or in some cases from results of programs having similar goals.

Evaluation Design

Of all the models, this one has the most difficulty establishing a suitable comparison standard. Programs making claims of intermediate improvement in practices and procedures may fall into one of two categories:

- The competitive practice. Certain services provided by education-related institutions are fairly standard; they have traditionally been provided, and probably will always be needed. In schools these include basic instructional and support functions. In libraries they include things like the circulation of books and the provision of reference services. In museums, planetariums, zoos, or aquariums, they include the dissemination of knowledge about natural phenomena and cultural or historical artifacts. A given program may have better methods for these standard activities than do most programs in similar institutions. Compared to others, the program may result in greater efficiency, increased use, or lower costs.

- The unique practice. This may be a program or practice that is being reported for the first time, that addresses different goals and claims from any seen previously. It may result from a new technology, an attempt to serve an unserved population, or the introduction of new knowledge. The changes produced by such a program must be

assessed strictly on their own merits, for there are no data on similar programs with which to make comparisons. Nonetheless, the program may be a successful and highly valuable innovation that merits widespread adoption.

In the case of a "competitive practice," it should be possible to compare a program's results not only with preexisting conditions, but also with the results of programs addressing similar goals in similar locations. Thus, appropriate standards of comparison are available and the evaluation task should be similar to that for other models.

In the case of a "unique practice," there may be no legitimate standard of comparison. The comparison to preexisting conditions should certainly be made; at least it establishes the existence of change. Program evaluators should be able to answer questions such as the following: Is a demand created? Of what size? How permanent? Who are the users? What are their comments? How do costs and usage compare?

The problem still remains that comparisons between a "unique practice" and a "do nothing" treatment are basically artificial, proving that something is better than nothing does not prove that it is worthwhile. For programs of the unique type, a sound evaluation design is important, but because it cannot offer a realistic standard of comparison, it provides far less support than usual for claims of effectiveness. The evaluator or program developer will have to make up for this deficiency by addressing the question of educational importance thoroughly and persuasively.

In the future, the process of panel review is likely to contribute to the development of standards for consideration of programs which claim to improve practices. As such programs come before the panel in growing numbers, a body of comparison data will emerge for certain types of programs. As claims are scrutinized and the body of knowledge grows, standards will be formulated, discussed, challenged, and adjusted, just as they have been for more traditional educational programs in the past.

Potential Panel Questions About Design:

In addition to questions raised for previous models, the following apply:

- Has the design used the most realistic no-treatment expectation available?
- If comparison was to other treatments, how similar were the situations?
- Does the design allow for pre-post assessment or time-series assessment over a time period sufficiently long to assure stability?
- Has attention been paid to assessment of implementation? Has the population in question been exposed to the treatment in a uniform way?
- Is a variety of measures and comparisons used?
- If sampling was used, how representative is each sample of the larger group?

- Does the design allow assessment of progress toward long-term academic goals?

Instruments, Procedures, and Data Collection

As with other intermediate models, gathering information from a variety of measures and sources is helpful. Unobtrusive measures such as institutional records are a likely choice, but care should be exercised to make sure that they are appropriate to the claims being made, and that recordkeeping procedures remain uniform throughout the course of the study.

Questionnaires and surveys of the target population are sometimes used, but unless appropriate sampling procedures are followed, these may yield biased results. For example, questionnaire returns from a convenience sample (*i.e.*, volunteers) are likely to show a very different response pattern from that of a more representative sample. The people most likely to complete the form are those who are very enthusiastic about the service, or those who have a complaint about it — not necessarily the most typical users. As always, the validity and reliability of such instruments should be discussed.

Potential Panel Questions About Instruments, Procedures, and Data Collection:

In addition to questions raised under previous models, the following apply:

- Is the information collected valid for the claims made?
- If existing records are used, what safeguards exist to ensure accurate completion and maintenance of recorded information?
- Is information collected from records complete?
- Does the information reported reflect all aspects of the treatment?
- Are the reporting units reasonable for purposes of comparison with existing standards?
- If self-reporting measures were used, what was done to promote objectivity? What was done to deal with nonresponse?

Analysis and Discussion of Results

For the "competitive practice," requirements for the analysis and discussion of results are similar to those for other models. For the "unique practice," as we have seen, there may be special problems in demonstrating educational significance. These problems arise from the difficulty of determining effect size or substantive value when a realistic standard of comparison cannot be identified.

Even in the absence of comparison data, the panel can, as it has for innovative educational programs in the past, develop a basis for judgment. Program cost may be considered. A demonstrated change may be produced at a low enough cost to be judged worthwhile. Another factor is the potential demand for such a program based on developing needs in a particular content area or target population. Demand may be related to emerging government or social priorities. Perhaps many other communities

could benefit from a similar program, and could afford it. Ease of adoption could also be important in the panel's view. If potential adopter sites are capable of reproducing the program, and if ample documentation is available, then a stronger case for approval might be presented.

As with all intermediate models, a clear presentation should be made of the hypothesized link between the immediate program outcomes and student achievement.

Potential Panel Questions About Analysis And Discussion Of Results:

In addition to questions raised for previous claim types, the following apply:

- Do the changes or improvements compare favorably to standard practices in similar institutions?
- Is there a clear link between potential student achievement and the attainment of immediate goals?
- Are the ultimate goals worthy ones?
- Is the demonstrated change worthwhile in terms of cost? Are the time savings worthwhile?
- Are there any unintended negative effects from the program? For example, are time or resources taken away from other areas?
- Are there rival hypotheses that could account for the observed effects?
- Is the scope of the change significant? Are other institutions likely to be interested in adopting the treatment?

Special Note: Schoolwide or Systemwide Change

The types of claims described in Chapter 5 are appropriate for target groups, project, school and systemwide levels but special issues should be considered with large-scale interventions. The special evaluation demands of projects designed to change an entire school or school system are usually related to the holistic nature of long-term institutional improvements. Typically such reforms are designed to support effective instruction across several disciplines, rather than to focus on a particular curriculum component. Examples might be attempts to raise overall school achievement, systems to enhance the efficiency of instruction, or plans to enhance teachers' effectiveness.

A submission claiming systemwide change should clearly present the organizing principles which account for the overall effect and whose adoption would enable other schools to achieve similar results. It is the special burden of projects of this scope to demonstrate that the outcomes observed are attributable to the organizing principles and not to particular circumstances, context, or staff. For this reason claims of systemwide change are usually based on observations over several years. Because of the scope of the change measured, it is wise to present data from a variety of measures in order to support the conclusions and to supplement data with narrative accounts that convey how a particular approach successfully united various elements to achieve the synergistic effect.

In building a case for a systemwide effect, it is necessary to present enough information, which is sufficiently representative of the system, over an adequate period of time. While this challenge seems major, systemwide projects sometimes offer the opportunity to "piggyback" upon routine data collection or to use historical data. School or district achievement test scores may be available over a number of years; data from unobtrusive measures in the form of regular school records might be useful. In addition, such data can be supplemented by information from surveys, interviews, or observations. Appropriate sampling procedures should be implemented when data is collected from a subset of the total population.

As always, the results must be compared to an appropriate standard or baseline. The standard can be internal, for example, longitudinal evidence showing the coincidence of reform implementation (or varying degrees of implementation) with gains in student achievement. Alternatively, the design could be based on cross-sectional comparisons of results with schools of similar socioeconomic composition and financial resources. Because of the multi-faceted nature of the treatment and the large scope of the effects, there are many opportunities for confounding elements. A coherent and logical presentation of a few basic changes that constitute a meaningful reform, and their logical relationship to significant outcomes, is essential. There is always the danger of mistaking trivial shifts in resources and programmatic emphasis for substantive changes in the schooling process.

A major challenge in the evaluation of systemwide projects is control of the many possible alternative causes that may exist in the whole system. It is especially important to be wary of changes in outcome variables that may result from shifts in the socioeconomic characteristics of the school population.

PEP Questions About Systemwide Change

In reviewing projects making claims of systemwide change, the panel will take these questions into consideration:

- If the baseline is external, are the pre-program similarities of the treatment and comparison schools or systems documented for all educationally relevant variables in addition to those directly addressed by the program (*e.g.*, racial, ethnic, socioeconomic, institution size, and resources)?
- If the baseline is internal, are the pre- and post-program conditions documented by sufficient data to overcome normal fluctuations and demonstrate trends convincingly?
- If samples from the entire population are used, are they representative? Are they large enough to generalize with confidence to the population as a whole?
- Are the students performing or participating at greater-than-predicted levels after controlling for other student variables including socioeconomic status?
- Are the effects sustained or do they represent gains in a single year which diminish over time, or gains in a single grade which diminish in later grades?
- Is there a plausible relationship between the nature of the changes made and long-term or large-scale effects being claimed?
- Does the change represent *more* than shifts in programmatic emphases and resources which were perhaps previously neglected? Have there been shifts in staff or allocated time?
- Did the implementation coincide with other changes in the school's organization or resources which are not viewed as part of the systemwide reform program? Could these other changes have contributed to the observed effects?
- Have there been changes in the population of the school or district? Is evidence presented to support the stability of conditions?
- Have there been changes in measures or recordkeeping procedures over the course of the study that could affect the nature or size of outcomes? (This issue is a particularly important question when making use of school or district data collection not done specifically for the program evaluation.)
- Is it possible that changes simply reflect developing trends in the larger educational milieu, *e.g.*, the influence of new state or federal requirements or changing social values?

Evaluators must accomplish a difficult task—to make a clear and logical presentation of various types of data showing the effects of several discrete elements that comprise a systemwide reform, and to relate these effects to long-term, large-scale outcomes.

CHAPTER 6

SUGGESTED FORMAT FOR PROJECT SUBMISSION

Over time, PEP has developed a preferred format for presentation of evidence. The format suggested here allows the development of a logical argument in a succinct fashion. Note the page limitation of 15 pages. Projects are expected to adhere to that requirement and to follow the suggested format to the extent possible.

Abstract

The abstract should be a 1-page description (200–300 words) of the program which provides a concise statement of concrete, observable outcomes. The abstract should briefly describe the following aspects of the program:

- goals
- purposes and needs addressed
- method of operation
- audience
- claim(s)

The abstract serves as a cover page to the 15-page submittal.

Basic Information

Basic information should be approximately 1 page.

A. Project Title
Location
Contact Person

Give the title of the project (including any acronym or abbreviation), the name of the applicant agency, and the address and a daytime telephone number of a contact person within the applicant agency.

B. Original Developer
Applicant Agency

Provide the name(s) and title(s) of those who originally developed the program. Describe the mission of the applicant agency and its legal status (*e.g.*, school district, nonprofit corporation).

C. Years of Project

Date(s) developed
Date(s) operated
Date(s) evaluated
Date(s) disseminated (if prior dissemination has occurred)

D. Source(s) and Level(s) of Development and Dissemination Funding

Federal	State	Local	Other	Total

List funding sources for the project and the amounts, by year.

Description of Program

Describe the program in approximately 5–6 pages.

A. Goals

Provide a clear and concise statement of the program goals. Include only those goals that relate directly to claims of effectiveness. In the case of evaluation models designed to meet intermediate objectives (*e.g.*, changing teaching strategies), make the link to the ultimate educational purpose of the program.

B. Purposes and Needs Addressed

Describe the specific needs the program was designed to address. Needs should be linked to the target audience and special features of the treatment.

C. Intended Audience

Identify the relevant demographic and educational characteristics of the population for which the intervention is designed. Use descriptors that pertain to grade level, content area, ability level, and achievement as they apply.

D. Background, Foundation, and Theoretical Framework

Discuss briefly the history of how and why the program was developed. Present the theoretical or empirical framework upon which the program is based. Include literature citations or research summaries as appropriate.

E. Features: How the Program Operates

Provide a complete description of how the program actually operates, identifying all features critical to implementation. Include the following topics as they apply to the project:

- scope (Does the project supplement or replace an existing program, or is it a component of a larger program?)
- curriculum and instructional approach
- learner activities
- learning materials
- staff activities and staffing patterns
- staff development activities
- management activities
- monitoring and evaluation procedures

F. Significance of Program Design as Compared to Similar Programs

Describe the features of the program that distinguish it from similar programs. Discuss ways in which the program addresses special problems. Note innovative or unique features. Tell how the program responds to state-of-the-art standards in its field.

Potential for Replication

Describe the potential for replication in approximately 2–3 pages.

A. Settings and Participants (Development and Evaluation Sites)

Briefly describe the community(ies) where the intervention was developed or field tested. Socioeconomic, ethnic, and geographic descriptions are appropriate. Also provide a brief description of the type of educational agency or school district(s) involved in the project. Include factors such as enrollment, ethnic composition, and general achievement level of students.

Specify the relevant demographic and educational characteristics of the population involved in developing or field testing the intervention. This population may differ from the intended target group. For instance, the intervention may be designed for students in grades K–12, whereas the project was used and tested only on students in grades 3, 5, 7, 9, and 11. In such a case, the latter group of students should be described here.

B. Replicable Components and Documentation

Indicate which aspects of the program are appropriate for dissemination to other sites. If the program has already developed support materials for dissemination (*e.g.*, teacher manuals), indicate the type of documentation available.

C. User Requirements

Describe the minimal requirements necessary for implementing the project in another site (*e.g.*, special staff, facilities, staff training time).

D. Costs (for Implementation and Operation)

Present a brief explanation of the recurring and nonrecurring costs associated with adopting the project. Costs such as personnel costs, special equipment, materials and supplies that are necessary for installing and/or maintaining the program at an adopting site should be discussed. Costs associated with the development of the original program should be excluded from this discussion.

Cost Table

Item	Start-up	Operation
Personnel		
Training		
Equipment		
Materials and Supplies		
Other		
Total Cost		
Cost Per Student/User		

Evidence

Provide evidence in approximately 6–8 pages.

A. Claim(s) Statement(s)

Programs should identify the specific PEP evaluation model(s) which best represent(s) its claims.

Succinctly state the major accomplishments of the project. Include a brief description of the type of evidence used to support the claim statement(s) (*e.g.*, test scores) and the nature of change that was demonstrated (*e.g.*, student achievement).

Generally, a claim statement includes:

- the target group for which results are available;
- the nature of the change;
- the process used for measuring gains; and
- the standards by which to judge gains as significant.

A clear claim statement is critical because the panel judges the adequacy of evidence based on the claim. Further, the claim identifies the project objectives/outcomes that will be approved for dissemination (*i.e.*, only those objectives/outcomes which are reflected in the claim

statement(s) and supported by convincing evidence will be approved). Sample claim statements are found in the description of each model in the previous chapter.

B. Description Of Methodology for Each Claim

1. Design

An evaluation design usually addresses three factors:

- the timing of data collection (*e.g.*, pre- and post-tests or different points in a time series)
- the groups involved (*e.g.*, a group receiving the program and a comparison group receiving an alternate program)
- the way in which a standard of comparison will be determined (*e.g.*, treatment group's gain or change will be compared to national or state benchmarks).

Describe the type of design used for each claim and the reason for the choice. Address any assumptions or problems inherent in the research design that was used.

2. Sample

The discussion of sampling procedures should answer four questions:

- Who participated in the study?
- How was the sample selected?
- How many participants were included in the final sample?
- How representative is the sample of the target population and program participants as a whole?

3. Instruments and Procedures

This section should describe the instruments and/or procedures and how each assessment technique relates to the outcomes. Provide sufficient information so that a judgment can be made about the technical strength and appropriate use of the measure (*e.g.*, validity, reliability, levels, subscales).

It is especially important to describe validity and reliability procedures for project-developed instruments; in such cases, the procedures for instrument development and field-testing also should be explained.

4. Data Collection

Describe the procedures used to select and train testers and the actual strategies used to assure quality control during data collection. Indicate the periods of data collection, the persons responsible for supervising

the data collection, and scoring and data summary procedures. It is especially important to describe in detail the data collection and quality control procedures for qualitative evaluations.

5. Data Analysis

If data are quantitative in nature, indicate the statistical technique(s) and levels of significance used in the analysis; levels of significance are usually set at the .01 or .05 probability level. Specify the criterion used in establishing effect size; generally this is presented as some proportion of a standard deviation.

If data are qualitative in nature, describe the procedures used to code and categorize or reduce information for summary purposes. Describe ways in which linkages were made across data elements to draw and verify conclusions.

C. Description of Results For Each Claim

Present detailed results of analyses in table or chart form, if appropriate. Sufficient detail should be provided for the reader to check conclusions independently. Also summarize the results for the claims in narrative form, relating the specific outcomes to the accomplishment of goals.

D. Summary of Supplementary Evidence for Each Claim

Provide additional evidence that supports the main claim, including anecdotal information, perceptions of quality, and levels of satisfaction. Supplementary evidence can also provide evidence of generalizability.

E. Interpretation and Discussion of Results

1. Relationship Between Effect and Treatment

Summarize the results of all data related to the claim(s) of effectiveness. Link the results to specific features of the program design.

2. Control of Rival Hypotheses

Provide evidence of program attribution, *i.e.*, evidence which suggests that the effects can be attributed to the program and not to some other equally plausible factor. As appropriate to the design, show how the following alternative explanations can be eliminated from consideration: maturation, other treatments, historical factors, statistical regression, attrition, differential selection of groups, and testing. (Note: Sound evaluation design can control most rival hypotheses; however, other data may be used to show attribution of effects.)

F. Educational Significance of Results

1. Relationship of Results to Needs

Demonstrate how the obtained results are important: how results meet the needs for which the project was designed. Establish the importance of the needs and demonstrate that the results are large enough and powerful enough to be viewed as important.

2. Comparison of Results to Results from Other Programs

Compare the results with results of similar projects or national or statewide initiatives, if appropriate.

BIBLIOGRAPHY OF USEFUL EVALUATION REFERENCES

There are a large number of books as well as articles in professional journals that address issues of concern to school-based personnel who are interested in applying for approval by PEP. Some of these are quite technical while others are more general in nature. The references listed below are offered as a sample of what is available; the list is not intended to be exhaustive.

For convenience, the information in this bibliography has been divided into three sections: General Statistics and Research; General Evaluation Methods and Instrumentation; and Qualitative Techniques and Case Studies.

GENERAL STATISTICS AND RESEARCH

Berkowitz, L. & Donnerstein, E. "External validity is more than skin deep." *American Psychologist*, 1982, 37, 245-257.

Campbell, D.T. & Stanley, J.C. *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally, 1966.

Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis of the behavioral sciences.* Hillsdale, NJ: Erlbaum Associates, 1983.

Edwards, A. *Experimental designs in psychological research.* New York, NY: Holt, Rinehart & Winston, 1972.

Ferguson, G.A. *Statistical analysis in psychology and education.* New York, NY: McGraw-Hill, 1981.

Joint Committee on Standards for Educational Evaluation. *Standards for evaluation of educational programs, projects, and materials.* New York, NY: McGraw-Hill, 1981.

Hinkle, D.E., Wiersma, W., & Jurs, S. G. *Applied statistics for the behavioral sciences.* Chicago, IL: Rand McNally, 1979.

Mehrens, W.A. & Lehmann, I.J. *Using standardized tests in education.* New York, NY: Longman Inc., 1987.

Popham, W.J. *Modern educational measurement.* Englewood Cliffs, NJ: Prentice-Hall, Inc., 1981.

Wiersma, W. *Research methods in education: an introduction.* Itasca, IL: F.E. Peacock Publishers, Inc., 1975.

Winer, B.J. *Statistical principles in experimental design.* New York, NY: McGraw-Hill, 1971.

Wood, R. *Measurement and assessment in education and psychology.* Philadelphia, PA: Falmer Press, 1987.

GENERAL EVALUATION METHODS AND INSTRUMENTATION

Berk, R.A. (ed.) *A guide to criterion-referenced test construction.* Baltimore, MD: The Johns Hopkins University Press, 1984.

Berk, R.A. (ed.) *Educational evaluation methodology: the state of the art.* Baltimore, MD: The Johns Hopkins University Press, 1981.

Borich, G.D. (ed.) *Evaluating educational programs and products.* Englewood Cliffs, NJ: Educational Technology Publications, 1974.

Brinkerhoff, R.O., et al. *Program evaluation: a practitioner's guide for trainers and educators.* Boston, MA: Kluwer-Nijhoff Publishing, 1983.

Cook, T.D., & Campbell, D.T. *Quasi-experimentation: design and analysis issues for field settings.* Chicago, IL: Rand McNally, 1979.

Cook T, & C. Reichardt (eds.), *Qualitative and quantitative methods in evaluation research.* Beverly Hills, CA: Sage Publications, 1979.

Cooley, W.W. "Explanatory observational studies." *Educational Researcher,* 1978, 7(9), 9-15.

Cronbach, L.J. *Designing evaluations of educational and social programs.* San Francisco, CA: Jossey-Bass Publishers, 1982.

Cronbach, L. et al. *Toward reform of program evaluation.* San Francisco, CA: Jossey-Bass, 1980.

Datta, L.E. (ed.) *Evaluation in change: meeting new government needs.* Beverly Hills, CA: Sage Publications, 1981.

Fink, A. & Kosecoff, J. *How to evaluate education programs: A compilation of ideas and methods that work.* Washington, DC: Capitol Publications, Inc., 1980.

Guba, E.G. & Lincoln, Y.S. *Effective evaluation.* San Francisco, CA: Jossey-Bass, 1981.

Henerson, M.E., et al. *How to measure attitudes.* Beverly Hills, CA: Sage Publications, 1978.

House, E.R. (ed.) *School evaluation: the politics and process.* Berkeley, CA: McCutchan, 1973.

Kirusek, T.J. & Lund, S.H. "Process and outcome measurements using goal attainment scaling." In G.V. Glass (ed.), *Evaluation studies review annual.* Beverly Hills, CA: Sage Publications, 1976.

Madaus, G.F., Scriven, M., & Stufflebeam, D.L. *Evaluation models: viewpoints on educational and human services evaluation.* Boston, MA: Kluwer-Nijhoff Publications, 1983.

Morris, L.L. *Program evaluation kit.* Beverly Hills, CA: Sage Publications, 1978.

Morris, L.L. & Fitz-Gibbon, C.T. *How to present an evaluation report.* Beverly Hills, CA: Sage Publications, 1978.

Oppenheim, A.N. *Questionnaire design and attitude measurement.* New York, NY: Basic Books, 1966.

Popham, W.J. *An evaluation guidebook: a set of practical guidelines for the educational evaluator.* Los Angeles, CA: The Instructional Objectives Exchange, 1972.

Popham, W.J. (ed.) *Evaluation in education: current applications.* Berkeley, CA: McCutchan, 1974.

Rich, J. *Interviewing children and adolescents.* London: Macmillan, 1968.

Stake, R.E. "The case study method in social inquiry." Educational Researcher, 1978, 7(2), 5-8

Struening, E. & Guttentag, M. (eds.) *Handbook of evaluation research.* Beverly Hills, CA: Sage Publications, 1975.

Tuckerman, B.W. *Evaluating instructional programs.* Boston, MA: Allyn & Bacon, Inc., 1979.

U.S. General Accounting Office. *Evaluation and analysis to support decisionmaking.* Washington, DC: U.S. GAO, 1976.

Weiss, C.H. *Evaluation research: methods of assessing program effectiveness.* Englewood Cliffs, NJ: Prentice-Hall, Inc., 1972.

Worthen, B.R. & Sanders, J.R. *Educational evaluation: Theory and practice.* Belmont, CA: Wadsworth Publishing Company, Inc., 1973.

Worthen, B.R. & White, K.R. *Evaluating educational and social programs: guidelines for proposal review, on-site evaluation, evaluation contracts, and technical assistance.* Hingham, MA: Kluwer-Nijhoff Publishing, 1987.

QUALITATIVE TECHNIQUES AND CASE STUDIES

Becker, H.S. & Greer, B. "Participant observation: The analysis of qualitative field data." In R.N. Adams & J.J. Priess (eds.), *Human organization research.* Homewood, IL: Dorsey Press, Inc., 1960.

Bogdan, R. & Biklen, S.K. *Qualitative research for education: an introduction to theory and methods.* Boston, MA: Allyn and Bacon, 1982.

Bogdan, R. & Taylor, S. *Introduction to qualitative research methods: A phenomenological approach to the social sciences.* New York, NY: John Wiley & Sons, 1975.

Cronbach, L.J. "Beyond the two disciplines of scientific psychology." *American Psychologist, 30,* 116-127, 1975.

Datta, Lois-ellin. *Case study evaluations.* Washington, DC: General Accounting Office, 1987.

Denny T. *Some still do: River Acres, Texas.* Urbana-Champagne, IL: Center for Instructional Research and Curriculum Evaluation, University of Illinois, 1977.

Douglas, J.D. *Investigative social research: individual and team field research.* Beverly Hills, CA: Sage Publications, 1976.

Erickson, F. "Qualitative methods in research on teaching." In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 119-161). New York, NY: Macmillan, 1986.

Erickson, F. "Some approaches to inquiry in school-community ethnography." *Anthropology and Education Quarterly, 8,* 58-69, 1977.

Fetterman, D.M. & Pitman, M.A. (eds.) *Educational evaluation: ethnography in theory, practice, and politics.* Beverly Hills, CA: Sage Publications, 1986.

Goetz, J., & LeCompte, M. *Ethnography and qualitative design in educational research.* New York, NY: Academic Press, 1984.

Guba, E.G. "Naturalistic evaluation." In D.S. Cordrey, H.S. Bloom, & R.S. Light (eds.) *Evaluation practice in review.* New Directions for Program Evaluation, No. 34. San Francisco, CA: Jossey-Bass, 1987.

Guba, E.G. *Toward a methodology of naturalistic inquiry in educational evaluation.* CSE Monograph Series in Education, Vol. 8. Los Angeles, CA: Center for the Study of Evaluation, 1978.

Guba, E.G. "What have we learned about naturalistic evaluation?" *Evaluation Practice, 8,* (1), 23-43, 1987.

House, E.R. *Evaluating with validity.* Beverly Hills, CA: Sage Publications, 1980.

Jacob, E. "Clarifying qualitative research: A focus on tradition." *Educational Researcher, 17,* (1), 16-24, 1988.

Lincoln, Y.S. & Guba, E.G. *Naturalistic inquiry.* Beverly Hills, CA: Sage Publications, 1985.

Miles, M.B. & Huberman, A.M. *Qualitative data analysis: a sourcebook of new methods.* Beverly Hills, CA: Sage Publications, 1984.

Patton, M.Q. *Qualitative evaluation methods.* Beverly Hills, CA: Sage Publications, 1984.

Richer, S." School effects: The case for grounded theory." *Sociology of Education,* 1975, *48,* 383-399.

Smith, L.M. & Geoffrey, W. *The complexities of an urban classroom: an analysis toward a general theory of teaching.* New York, NY: Holt, Rinehart & Winston, Inc., 1968.

Stake, R. & Gjerde, C. "An evaluation of TCITY, the Twin City Institute for Talented Youth." In Richard H.P. Kraft, et al (eds.), *Four evaluation examples: anthropological, economic, narrative and portrayal.* AERA Monograph Series on Curriculum Evaluation, Chicago, IL: Rand McNally College Publishing, 1973.

Von Maanen, J. (ed.) *Qualitative methodology.* Beverly Hills, CA: Sage Publications, 1983.

Webb, E.J. et al. *Unobtrusive measures: nonreactive research in the social sciences.* Chicago, IL: Rand McNally & Co., 1966.

Whyte, W.F. *Learning from the field: a guide from experience.* Newbury Park, CA: Sage Publications, 1984.

Wold, H. "Causal inferences from observational data: A review of ends and means." *Royal Statistical Society Journal,* 1956, Series A, *119,* 28-50.

Woods, P. *Inside schools: ethnography in educational research.* Toronto, Canada: Routledge & Kegan Paul, 1986.

Yin, R. *Case study research: design and methods.* Beverly Hills, CA: Sage Publications, 1984.